# Capstone Project Proposal

The client for my first Capstone Project is Lending Club though any financial institution would benefit from such an analysis. I will be analyzing the loan data to create an algorithm, which would predict if a particular loan is good, neutral or bad (will default) to put it simplistically.

Lending has an inherent risk to it and the goal is to minimize that risk. Through this project, I can answer the ultimate question of whom should a loan be given and where the bank needs to be more cautious which would help Lending Club take necessary actions when processing a loan application.

I will be using the loan data for Lending Club that is available on kaggle. The below link can be used to access the data set.

Data link: <https://www.kaggle.com/wendykan/lending-club-loan-data>

This problem can be approached in a number of ways. I am planning to approach it like this:

1. Do some preliminary exploratory data analysis
   1. Understanding the structure of my data, the data types of columns and the number of observations.
   2. Ensuring there are no null values in the data/columns, which are being used.
   3. This would include doing some counts and running some comparisons, understanding correlations, if any.
2. Considering this is a supervised learning problem, I think it is classification problem and so I would use logistic regression.
3. Once my data is cleaned, I will choose the features, which will help me understand and predict if the loan in question is good or bad.
4. Once I have my features selected, I will split my data into testing and training set.
5. I would create a model which would then classify the loans in one of the two categories:
   1. Good
   2. Bad
6. I will then fit the model on my training data making sure I use cross validation.
7. Once I have the model trained on my training data, I will test it against the testing data and see how it performed by calculating the calculation report instead of accuracy score as that gives you a bigger picture of how is your model performing.
8. It’s a good practice to calculate the calculation report for both test and training data set to get a deeper insight.
9. Once I have a model ready, I can present my insights with the help of visualization and draw conclusions.

My deliverables for this capstone project will include:

1. Code to reproduce my analysis
2. A document, which further summarizes what was done and the outcome in a visual format wherever possible.